Interpretable and predictive models to harness the life science data revolution

Joshua P. Jahner[a,1], C. Alex Buerkle[a], Dustin G. Gannon[a], Eliza M. Grames[b,c], S. Eryn McFarlane[a,d], Andrew Siefert[a], Katherine L. Bell[b], Victoria L. DeLeo[a], Matthew L. Forister[b], Joshua G. Harrison[a], Daniel C. Laughlin[a], Amy C. Patterson[a], Breanna F. Powers[a,e], Chhaya M. Werner[a,f], Isabella A. Oleksy[g,h]

[a]Department of Botany, University of Wyoming, Laramie, WY 82071, USA

[b]Department of Biology, University of Nevada, Reno, NV, 89557 USA

[c]Department of Biological Sciences, Binghamton University, Binghamton, NY 13902 USA

[d]Department of Biology, York University, Toronto, ON M3J 1P3, Canada

[e]School of Forestry, Northern Arizona University, Flagstaff, AZ 86011, USA

[f]Department of Environmental Science, Policy & Sustainability, Southern Oregon University, Ashland, OR 97520, USA

[g]Department of Zoology and Physiology, University of Wyoming, Laramie, WY 82071, USA

[h]Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO 80303, USA

[1]To whom correspondence may be addressed. Email: jpjahner@gmail.com

Keywords: irreducible error, reducible error, simulation, sparse modeling, statistical learning

Abstract

The proliferation of biological data with large numbers of samples and many dimensions is kindling hope that life scientists will be able to fit statistical and machine learning models that are highly predictive and interpretable. However, large biological data sets are commonly burdened with an inherent trade-off: in-sample prediction will improve as additional predictors are included in the model, but this may come at the cost of poor predictive accuracy and limited generalizability for future or unsampled observations (i.e., out-of-sample prediction). To confront this problem of overfitting, sparse modeling methods can correctly place low weight on unimportant predictor variables, thereby narrowing in on the key predictors associated with variation in the response. We competed nine methods to quantify their performance in variable selection, estimation, and prediction using simulated data sets with different sample sizes, numbers of predictors, and strengths of effects. Overfitting was typical for many methods and all simulation scenarios, as demonstrated by large in-sample R^2 and low out-of-sample R^2. Despite this, in-sample and out-of-sample prediction converged on the true predictive target for simulations with more observations, larger effects of causal variables, and fewer predictors. Analyses of simulations showed that accurate variable selection to support process-based understanding will be unattainable for many realistic sampling schemes. We use our analyses to characterize data attributes in which statistical learning is possible, and illustrate how some sparse methods can be used to achieve predictive accuracy while mitigating and learning the extent of potential overfitting.

Introduction

The life sciences have seen dramatic growth in data due to technological advances for automation and high-throughput sampling (e.g., water or air sampling, Porter *et al.* 2009; satellite imagery, Ustin & Middleton 2021, Cavender-Bares *et al.* 2022; DNA sequencing, Halldorsson *et al.* 2022, Rubinacci *et al.* 2023; and GPS telemetry, Wilmers *et al.* 2015, Gigliotti *et al.* 2022). While large data sets have the potential to greatly improve our understanding of complex systems, they also pose considerable challenges for data analysis and incorporation into formal process models. For example, geneticists have sought to predict individual traits (e.g., human height) using data from $>$ 600,000 regions of the genome for 500,000 individuals, but models including all predictors (genetic loci) can yield many spurious associations and poor predictive performance for unsampled individuals and populations (Lello *et al.* 2018). Limited generalizability is common in parameter-rich models and results from overfitting, or the tendency for flexible models to fit too closely to the observed data such that idiosyncratic variation in the observed data is taken as pattern rather than noise (Hastie *et al.* 2015). Thus, the availability of big data with potentially many more covariates (P) than observations (N) may counter-intuitively lead to models with poor predictive performance outside the scope of the sample ("the curse of dimensionality", Altman & Krzywinski 2018). More broadly, we face the problem of how to realistically and intelligently constrain the flexibility of our models to capture potential general patterns while learning about genuine context-dependent effects (Weiss 2008).

We have made great strides in the life sciences in constructing highly predictive models based on computational modeling and machine learning to address the proliferation of large data sets. Machine learning complements standard methods for statistical modeling that make *a priori* choices of which predictor variables to include based on conceptual, process-based understanding. More generally, advances in machine learning have prompted a reevaluation of how we value models and what it means for a model to 'understand' something about the world (Mitchell & Krakauer 2023). Statistical models can be used to learn which predictors are associated with a response (i.e., variable

or feature selection), to generate accurate predictions about the sampled population (i.e., in-sample prediction), and to make generalizations about populations we have no prior information about (i.e., out-of-sample prediction). We value models that can reveal predictors that are associated with the generative processes leading to variation in the response, while also avoiding shortcut learning that can garner accurate predictions from tangentially related variables (shortcut learning can be problematic for both inference and out-of-sample prediction; Geirhos *et al.* 2020).

Life scientists would benefit from a direct comparison and evaluation of the prospects of different statistical learning methods, and from a greater clarity about critical issues in model evaluation, including overfitting, the extent to which process variance is recovered in model predictions, and the explanatory value of important predictor variables. Computational methods for sparse modeling might be particularly valuable approaches: these methods assume that most predictors have no causal relationship with the response and therefore only generate estimates for a subset of variables (Hastie *et al.* 2015). The hope is that selected predictors correspond to the key process variables that are causally linked to variation in the response, which should limit overfitting and improve predictive performance when generalizing to unsampled or future observations. It is an open question to what extent sparse models can maximize predictive performance and yield interpretable model outputs, particularly for data where P is much greater than N.

We compared the relative performance of several modeling methods by applying them to the same data sets with known, simulated causal relationships, of the type commonly encountered in the life sciences. Our 36 core simulation scenarios (100 simulated replicates each) differed in the number of observations (N = 50, 150, or 500), the number of covariates (P = 100, 1,000, 10,000, or 100,000; of which 10 were causal and directly influenced the response variable), and the effect size of causal predictors ($\boldsymbol{\beta}_{\text{causal}}$ = 0.1, 0.3, or 0.8; Table S1). Our statistical learning methods included penalized regression methods based on maximum likelihood (Ridge, Elastic Net, and LASSO) and Bayesian estimation (Bayesian LASSO [BLASSO], Horseshoe, Spike-and-slab, Sum of Single Effects [SuSiE], and Bayesian Sparse Linear Mixed Model [BSLMM]), and one commonly used machine learning method (Random Forest). When evaluating the strengths and weaknesses of different methods, we considered prediction (the accuracy of prediction of the response variable given the covariates, both for in-sample, training data and out-of-sample, test data) and inference (i.e., learning which variables are causally associated with variation in the response). While prediction and inference should be treated as complementary goals in statistical analyses (Breiman 2001b), it is worth noting they are not always associated, even if there is an expectation that strong inference would follow from accurate predictions (Wang *et al.* 2020b).

Results

A perfect model would: 1) identify only the ten truly causal predictors and accurately estimate their effect sizes; 2) accurately attribute the variation in the response that arises directly from the causal predictors (i.e., reducible error); and 3) disregard variation in the response arising from other unmeasured or stochastic processes (i.e., irreducible error; James *et al.* 2021). Across all simulations, the magnitude of reducible error was overwhelmingly associated with the effect size of causal predictors (R^2 between the known, additive effects of simulated causal predictors and the simulated response variable was $\approx$ 0.10, 0.47, and 0.86 with $\boldsymbol{\beta}_{\text{causal}}$ = 0.1, 0.3, and 0.8, respectively; Fig. S1). Reducible error was more variable among replicates when N or P were low.

The different methods varied greatly in their performance for variable selection and prediction. For one example data set (N = 150, P = 10,000, $\boldsymbol{\beta}_{\text{causal}}$ = 0.8; see Fig. 1), LASSO `monomvn` had

the greatest success at delineating between causal and non-causal predictors (true positive rate [TPR] = 0.9; true negative rate [TNR] = 0.998). While Random Forest also correctly identified nine of the ten causal predictors (TPR = 0.9), it implicated a large proportion of non-causal variants as being associated with the response (TNR = 0.123). In contrast, BSLMM was relatively successful at excluding non-causal predictors (TNR = 0.958), but could only identify half of the causal predictors (TPR = 0.5). For prediction, the true reducible error for the example data set was $R^2 = 0.832$, which served as the target for both in-sample and out-of-sample prediction (based on 500 observations not used to train the model). For LASSO `monomvn`, in-sample prediction was very close to the reducible error ($R^2 = 0.819$), which translated to the highest success for out-of-sample prediction ($R^2 = 0.749$). In-sample prediction exceeded the reducible error for BSLMM ($R^2 = 0.961$), and this overfitting led to reduced out-of-sample prediction ($R^2 = 0.622$) relative to LASSO `monomvn`. Random Forest suffered from poor predictive yield, with both in-sample ($R^2 = 0.084$) and out-of-sample ($R^2 = 0.341$) comparisons, falling far short of the reducible error. Overall, LASSO `monomvn` provided the best balance between variable selection and prediction for the example data set.

Overfitting was rampant across all scenarios, as evidenced by large in-sample R^2 and low out-of-sample R^2 (Fig. 2A and B). It was also common for models to recover only a fraction of the reducible error in out-of-sample prediction, particularly for simulations with larger P (Fig. 2B). The accuracy of in-sample and out-of-sample prediction converged towards the reducible error target for simulations with larger $\boldsymbol{\beta}_{\text{causal}}$ and N and smaller P (Fig. 3). Out-of-sample predictive performance was not necessarily associated with more accurate variable selection, as out-of-sample R^2 matched the reducible error even with low F_1 for some scenarios (Fig. S2). Variable selection was first assessed for methods that return truly sparse parameter estimates (i.e., $\boldsymbol{\beta} = 0$; BSLMM, Elastic Net, LASSO, Spike-and-slab) or importance values (Random Forest), and was generally poor except from when $\boldsymbol{\beta}_{\text{causal}}$ and N were high and P was low; Fig. 2C). When $\boldsymbol{\beta}_{\text{causal}}$ was low, a negative relationship between TPR and TNR emerged across methods, suggesting a trade-off between identifying causal predictors and excluding non-causal predictors (Fig. 4). Variable selection was also assessed based on posterior inclusion probabilities (PIPs) for four Bayesian methods (BLASSO, BSLMM, Horseshoe, SuSiE) using the example data set from Fig. 1. The use of a small PIP threshold of 0.05 (i.e., only predictors with PIP ≥ 0.05 are scored as positives) improved variable selection for BSLMM and SuSiE, whereas larger thresholds were needed to recover more limited gains for BLASSO and Horseshoe (Fig. S3). Parameter estimation was remarkably consistent across different analyses, and was instead most strongly influenced by data dimensionality: estimation was worse with greater $\boldsymbol{\beta}_{\text{causal}}$ and lower N and P (Fig. 2D). This pattern arose because most methods are worse at estimating predictors with $\boldsymbol{\beta} \neq 0$ than those with $\boldsymbol{\beta} = 0$, resulting in larger root mean square error (RMSE) when the proportion of causal to non-causal predictors was relatively large (i.e., when P was small) or when the effect size of causal predictors was very different from zero (i.e., when $\boldsymbol{\beta}_{\text{causal}}$ was large). Analysis of the 3,600 data sets completed in 2.49 CPU-years, with BLASSO and Horseshoe contributing 46.6% and 46.7% of the total run-time, respectively (Fig. 2E).

Discussion

High-throughput and automated data acquisition promises to yield valuable information about processes that generate variation. This promise is diminished in the common situation in the life sciences when sampling is of few individuals (N) and many potential covariates (P; e.g., expression levels of 10^4 genes, genomic polymorphisms at 10^8 sites, months of micrometeorological sensor

measurements at 10Hz). Our simulations highlight that the most consistent way to obtain highly predictive and explanatory models is to maximize the number of independent observations. While sparse modeling techniques allow the fitting of models in settings with more covariates than observations ($P > N$), they cannot rescue analyses based on small sample sizes, especially when P is large or when effect sizes are small relative to background levels of stochastic variation (Fig. 2). This means that for many typical life science analyses, variable selection will suffer from low precision and sensitivity, and prediction models will be overfit and have poor generalizability. In cases where sparse methods struggle with a low signal-to-noise ratio, other methods will also struggle ("the bet-on-sparsity principle"; Hastie *et al.* 2009), meaning such signals will only ever be detectable with more data, better sampling design, or both. Indeed, when we extended our simulations to have sample sizes of 1,000 or 10,000 observations, in-sample and out-of-sample R^2 converged to the maximum reducible error, and variable selection improved for most analyses (Fig. 5).

It is perhaps näive to use statistical learning for prediction without large training sets, particularly when causal effect sizes are small relative to variance from extraneous sources. The temptation to do so might stem from working with big data ($N \times P$), but not appreciating that all statistical approaches are expected to yield relatively poor out-of-sample prediction when N is small (e.g., < 500) and effect sizes are modest. Some of the most remarkable models in society, such as those for large language modeling (Zhao *et al.* 2023), natural voice recognition (Xiong *et al.* 2016), image segmentation (Kirillov *et al.* 2023), and board game algorithms (Silver *et al.* 2018), are typically trained on enormous sample sizes. For example, Tabak *et al.* (2019) trained a convolutional neural network with more than 3 million images to achieve more than 80% out-of-sample accuracy when detecting ungulates from camera trap imagery. We do believe there is a place for sparse methods in the life sciences when many observations (N) can be obtained (Fig. 5). Our simulations provide context for evaluating different dimensions of model quality and the comparison of model approaches suggests which methods will be most useful and when.

For most predictive contexts, the primary objective is to account for the reducible error in the data, as this is the variation in the response associated with generative processes (James *et al.* 2021). We were able to directly quantify the reducible error in our simulated data sets and easily identify cases of overfitting in which in-sample $R^2 >$ reducible error R^2 (Figs. 2A & 3). With empirical data, the true reducible error and prediction errors arising from model variance and bias will be unknown (James *et al.* 2021), but overfitting may be evident when in-sample R^2 exceeds out-of-sample R^2. It is worth emphasizing that in our simulations and analyses, we minimized the potential for model bias and underfitting by simulating data from simple additive generative processes that are mirrored in the statistical learning methods we used. In other words, we simulated the best case scenarios for explaining reducible error, and we still typically fell short.

To minimize errors in prediction, we can strive for large sample sizes of representative data for model training (i.e., homogeneous with the out-of-sample, test data). Additionally, on average, in-sample prediction accuracy cannot be less than out-of-sample prediction accuracy, and both will converge on the true reducible error with increasing $\boldsymbol{\beta}_{\text{causal}}$ and N and decreasing P (Fig. 3). Recovery of similar in-sample and out-of-sample R^2 is consistent with minimal overfitting, but could arise from model bias. Similar out-of-sample R^2 from multiple, genuinely different analysis methods would be consistent with having minimized prediction error given the information in the available data, but could still derive from underfit, biased models that account for only a fraction of the true reducible error (see results from Random Forest in the example data set; Fig. 1). It is worth noting that Random Forest always yielded in-sample R^2 roughly equal to out-of-sample R^2 (Figs. 2A,B & 3), as this is the only method that uses cross-validation as a default. The distinction between prediction errors that arise in-sample and out-of-sample (Fig. 3), and the strong

potential for overfitting, call into question model choice decisions that are very commonly made based on in-sample data alone without any cross-validation procedure (i.e., potentially choosing the most overfit model with little deference for out-of-sample prediction; see Fig. 3; Tredennick *et al.* 2021). Importantly, while cross-validation is critical for safeguarding against misleading model results, reduced R^2 values from cross-validated models could result in studies being less likely to be published or going into a lower impact journal, suggesting the need for a shift in how researchers evaluate prediction results in the context of cross-validation.

The conditions that allow for strong prediction, namely when $\boldsymbol{\beta}_{\text{causal}}$ and N are large and P is small, are the same in which variable selection is possible (Figs. 2C & 3), though reliable out-of-sample prediction did not necessarily depend on perfect variable selection (i.e., including all causal and excluding all non-causal predictors; Fig. S2). A variable selection trade-off emerged for the data sets in which variable selection was most difficult (e.g., $\boldsymbol{\beta}_{\text{causal}} = 0.1$), as evidenced by a negative relationship between true positive and true negative rates (Fig. 4). This result has broad implications because effect sizes are expected to be small and diffuse for many biological systems (e.g., in genetics, Boyle *et al.* 2017). Moreover, the consequences of different variable selection errors will have disparate repercussions in exploratory versus diagnostic settings, so researchers will need to weigh the costs and benefits of either identifying all causal predictors at the expense of including some false positives (e.g., when developing candidate gene lists from genome-wide association studies) or missing some causal predictors to ensure the absence of any false positives (e.g., when identifying biomarkers for disease detection). For the Bayesian methods that generate posterior inclusion probabilities (PIPs), the threshold for deciding whether or not to include a variable may vary across disciplines and fields. In evolutionary genetics, researchers may choose to only consider genetic loci that have a PIP > 0.1 (Lucas *et al.* 2018, McFarlane & Pemberton 2021), and this simple choice would have substantially improved variable selection for BSLMM and SuSiE, but not BLASSO or Horseshoe, for one example scenario (Fig. S3). Overall, accurate variable selection requires large numbers of observations (Fig. 5), perhaps even more so than prediction, as has been found previously in trait mapping and phenotypic prediction (Wray *et al.* 2013, Gompert *et al.* 2017).

One striking feature of our results was the absence of a single method that excelled at all modeling purposes, consistent with the "no free lunch theorem" for supervised learning (Wolpert 1996, Wolpert & Macready 1997). Trade-offs in model building have long been recognized (Levins 1966, James *et al.* 2021, Tredennick *et al.* 2021) and serve as an important reminder for researchers to wield methods that align with their research objectives. Consequently, it can be useful to simulate data and measure the correlation (and other measures of the relationship) of the response variable with process parameters ($\boldsymbol{\beta}_{\text{causal}}$) under relevant sample sizes, so as to gauge information about the expected reducible error. It may be the case that researchers will need to employ multiple, complementary statistical learning methods for questions involving both prediction and variable selection. A combined approach to model building could be particularly valuable, for example using a sparse method to identify a subset of candidate variables and following up with a more flexible method such as Random Forest for prediction. We emphasize that while the use of sparse methods cannot resolve logistical challenges surrounding data collection in the life sciences (there will always be data sets where P is much greater than N), the uptake of these methods is a path forward that can contribute to high-quality inference, explanatory models that capture key elements of data generating processes, and prediction with minimal error. Finally, we acknowledge that many of our key findings recapitulate concepts that are already well known by many statisticians (James *et al.* 2021). Our simulations and analyses illustrate a number of points that are not widely appreciated in applied statistics, including in the life sciences, and we hope this exercise will elevate awareness

of the promise and limitations of these tools for statistical learning.

Materials and methods

Description of simulated data

The simulations included 36 scenarios that considered three main factors in a fully crossed design: the number of observed samples (N = 50, 150, or 500), the number of predictors or features (P = 100, 1,000, 10,000, or 100,000), and the effect size of the ten causal predictors ($\boldsymbol{\beta}_{\text{causal}}$ = 0.1, 0.3, or 0.8; Table S1). To evaluate the potential benefits of even larger N, we simulated two additional scenarios in which N was 1,000 or 10,000, P was 1,000, and $\boldsymbol{\beta}_{\text{causal}}$ was 0.3. To thoroughly incorporate and evaluate variable outcomes among simulations, we obtained 100 replicate data sets for all scenarios. Each replicate data set consisted of N observations for training (i.e., variable selection and in-sample prediction) and an additional 500 observations for testing out-of-sample prediction.

For each replicate, we first created an observation $\times$ predictor $(N+500)\times P$ matrix $\boldsymbol{X}$ consisting of $P/50$ clusters of correlated predictors (50 per cluster). Each cluster of predictors was generated by taking $N + 500$ draws from a multivariate normal distribution with mean vector $\boldsymbol{\mu} = 0$ and covariance matrix $\boldsymbol{\Sigma}$. We generated covariance matrices using a spherical parameterization (Pinheiro & Bates 1996), which transforms a $P(P+1)/2$-dimension vector of unconstrained parameters $\boldsymbol{\theta}$ into a positive semi-definite covariance matrix $\boldsymbol{\Sigma}$. The goal of this approach was to create clusters of predictors with a range of correlation strengths, from strongly negatively to strongly positively correlated, a situation that is common in biological relationships and that presents a challenge for many modeling approaches. We found that drawing values of $\boldsymbol{\theta}$ from a uniform distribution between -1 and 1 produced sets of predictors with a range of correlation strengths. After generating clusters of predictors, we concatenated them to create the predictor matrix $\boldsymbol{X}$ and centered and scaled (mean = 0; sd = 1) the columns of predictors.

Next, we sampled a P-dimension vector of coefficients $\boldsymbol{\beta}$ representing the causal effects of the predictors on response variable $\boldsymbol{y}$. We randomly selected 10 predictors out of P to have a non-zero coefficient of $\boldsymbol{\beta}_{\text{causal}}$. The remaining values of $\boldsymbol{\beta}$ were set to zero. The response variable $\boldsymbol{y}$ was a linear, additive function of the product of the $\boldsymbol{\beta}$ coefficients and the P predictors, plus error or intercept term of $\boldsymbol{\epsilon}$, drawn from a standard normal distribution for each individual: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. For each data set, the reducible error was calculated as the proportion of variance in the response explained by a linear model using only the 10 causal predictors.

We made several decisions in simulating data that could influence our results and interpretation. For example, causal parameters in the simulated data sets were specified as simple linear effects, as opposed to non-linear or threshold effects that could be more or less difficult to identify for some methods. However, while linear approximations of non-linear processes introduce bias, they can often outperform more flexible non-linear or non-parametric approaches that introduce more variance, particularly for high-dimensional data (i.e., "the bias-variance trade-off"; James *et al.* 2021). Furthermore, we intentionally avoided the complexities of causal inference in the presence of confounding variables and interactions. Instead, for the purpose of learning, we studied a simplified system in which sparse effects could estimate causal effects. Finally, we have explored a fairly small and simple range of data attributes that might be encountered in the life sciences, and acknowledge that the consideration of other axes of variation will undoubtedly lead to new insights about how we can use modeling approaches to better understand the world.

Analyses

Each simulated data set was modeled using nine different methods. Eight of these are penalized regression methods using standard likelihood (LASSO, Tibshirani 1996; Ridge, Hoerl & Kennard 1970; Elastic Net, Zou & Hastie 2005) or Bayesian estimation (Bayesian LASSO [BLASSO], Park & Casella 2008; Horseshoe, Carvalho *et al.* 2010; Spike-and-slab, Ishwaran & Rao 2005; Bayesian sparse linear mixed model [BSLMM], Zhou *et al.* 2013; sum of single effects [SuSiE], Wang *et al.* 2020a). The final method, Random Forest (Breiman 2001a), served as a benchmark to compare other methods to and is a commonly used, highly flexible machine learning approach based on an ensemble of decision trees. All analyses were conducted in `R v4.2.2` (R Core Team 2023). Each data set was provided to the methods using the `Nextflow v22.10.4.5836` workflow description language (Di Tommaso *et al.* 2017) to distribute the work and aggregate the output in a computing cluster using `SLURM` (Yoo *et al.* 2003). We used implementations of Elastic Net, LASSO, and Ridge in the `glmnet v4.1-6` package (Friedman *et al.* 2010), of BLASSO, Horseshoe, and alternatives of LASSO and Ridge in the `monomvn v1.9-17` package (Gramacy 2023), of Spike-and-slab in the `spikeslab v1.1.6` package (Ishwaran *et al.* 2010), of SuSiE in the `susieR v0.12.27` package (Wang *et al.* 2020a), of Random Forest in the `randomForest v4.7-1.1` package (Liaw & Wiener 2002), and of BSLMM in the software `gemma v0.98.6` (Zhou *et al.* 2013). We used 'off-the-shelf', default settings for all analyses (e.g., we did not tune hyperparameters). BLASSO and Horseshoe were not performed for the large N scenarios ($N = 1{,}000$ or $10{,}000$) due to extremely long run times. All `R` scripts for creating the simulated data, running analyses, generating summary statistics, and making figures, as well as the `Nextflow` wrappers and config files are archived at Zenodo (doi:10.5281/zenodo.10815222).

To evaluate each model's potential utility for parameter estimation, variable selection, and prediction, we calculated several complementary summary statistics that were largely applicable across all of the methods. Metrics for BLASSO and Horseshoe were calculated two ways: model-averaged (ma) estimates are based on all samples from the reversible jump MCMC, whereas non-zero (nz) estimates use only samples in which the predictor and associated coefficient were included in the model. Parameter estimation was evaluated based on the root mean square error (RMSE) between estimated and actual parameter values ($\boldsymbol{\beta}$) for all analyses except Random Forest, which reports importance measures instead of estimates. Variable selection was first assessed for methods that can return true zeros for parameter estimates (BSLMM, Elastic Net, LASSO, Spike-and-slab) or importance measures (Random Forest). Predictors were assigned as positives ($\neq 0$) or negatives ($= 0$), and these classifications were used to calculate true positive rates (TPR; i.e., sensitivity), true negative rates (TNR; i.e., specificity), and F_1, which is the harmonic mean of precision (i.e., the fraction of selected predictors that are truly causal) and sensitivity: $\frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision}$. It is important to note that small values of F_1 (i.e., poor variable selection) can occur due to low TPR, low TNR, or both. Variable selection was also assessed based on posterior inclusion probabilities (PIPs) for four Bayesian methods (BLASSO, BSLMM, Horseshoe, SuSiE) using one example data set (scenario 24, replicate 1). A series of minimum PIP thresholds (i.e., predictors with PIP $\geq$ threshold are scored as positives) were evaluated to characterize potential effects on resulting F_1 values. In-sample and out-of-sample prediction was quantified using R^2 between the actual and predicted values of the response variable. In-sample prediction was based on the N observations used to train the model, whereas out-of-sample prediction was based on a separate set of 500 observations. Finally, we recorded the runtime required to fit each model to each data set.

Acknowledgments

All authors were supported by the Modelscape Consortium with funding from the National Science Foundation (OIA-2019528). We thank: members of the Modelscape Consortium for input during early discussions of this project; Chris Moore for fun conversations about the philosophy of model building; Dan Gibson and Loren Rieseberg for thoughtful feedback on earlier drafts. Analyses were performed using University of Wyoming's Advanced Research Computing Center and its Beartooth Computing Environment, Intel x86_64 cluster (https://doi.org/10.15786/M2FY47)

Author contributions

J.P.J., C.A.B., S.E.M., A.S., M.L.F., D.C.L., B.F.P., and I.A.O. designed research; A.S. created simulations; J.P.J., C.A.B., D.G.G., E.M.G., S.E.M., A.S., J.G.H., C.M.W., and I.A.O. performed analyses; C.A.B., M.L.F., and D.C.L. acquired funding; all authors contributed to writing and revision.

References

Altman N, Krzywinski M (2018) The curse(s) of dimensionality. *Nature Methods*, **15**, 399–400.

Boyle EA, Li YI, Pritchard JK (2017) An expanded view of complex traits: from polygenic to omnigenic. *Cell*, **169**, 1177–1186.

Breiman L (2001a) Random forests. *Machine Learning*, **45**, 5–32.

Breiman L (2001b) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, **16**, 199–231.

Carvalho BS, Louis TA, Irizarry RA (2010) Quantifying uncertainty in genotype calls. *Bioinformatics*, **26**, 242–249.

Cavender-Bares J, Schneider FD, Santos MJ, *et al.* (2022) Integrating remote sensing with ecology and evolution to advance biodiversity conservation. *Nature Ecology & Evolution*, **6**, 506–519.

Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C (2017) Nextflow enables reproducible computational workflows. *Nature Biotechnology*, **35**, 316–319.

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1.

Geirhos R, Jacobsen JH, Michaelis C, *et al.* (2020) Shortcut learning in deep neural networks. *Nature Machine Intelligence*, **2**, 665–673.

Gigliotti LC, Xu W, Zuckerman GR, *et al.* (2022) Wildlife migrations highlight importance of both private lands and protected areas in the Greater Yellowstone Ecosystem. *Biological Conservation*, **275**, 109752.

Gompert Z, Egan SP, Barrett RD, Feder JL, Nosil P (2017) Multilocus approaches for the measurement of selection on correlated genetic loci. *Molecular ecology*, **26**, 365–382.

Gramacy RB (2023) *monomvn: Estimation for MVN and Student-t Data with Monotone Missingness*. R package version 1.9-17.

Halldorsson BV, Eggertsson HP, Moore KHS, *et al.* (2022) The sequences of 150,119 genomes in the UK Biobank. *Nature*, **607**, 732–740.

Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer.

Hastie T, Tibshirani R, Wainwright M (2015) *Statistical learning with sparsity*. Chapman & Hall.

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Ishwaran H, Kogalur UB, Rao JS (2010) spikeslab: Prediction and variable selection using spike and slab regression. *R Journal*, **2**, 68–73.

Ishwaran H, Rao JS (2005) Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, **33**, 730–773.

James G, Witten D, Hastie T, Tibshirani R (2021) *An introduction to statistical learning with applications in R*. 2nd edn., Springer.

Kirillov A, Mintun E, Ravi N, *et al.* (2023) Segment anything. *arXiv preprint arXiv:2304.02643*.

Lello L, Avery SG, Tellier L, Vazquez AI, de los Campos G, Hsu SDH (2018) Accurate genomic prediction of human height. *Genetics*, **210**, 477–497.

Levins R (1966) The strategy of model building in population biology. *American Scientist*, **54**, 421–431.

Liaw A, Wiener M (2002) Classification and regression by randomForest. *R news*, **2**, 18–22.

Lucas LK, Nice CC, Gompert Z (2018) Genetic constraints on wing pattern variation in *Lycaeides* butterflies: A case study on mapping complex, multifaceted traits in structured populations. *Molecular Ecology Resources*, **18**, 892–907.

McFarlane SE, Pemberton JM (2021) Admixture mapping reveals loci for carcass mass in red deer x sika hybrids in Kintyre, Scotland. *G3*, **11**, jkab274.

Mitchell M, Krakauer DC (2023) The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, **120**, e2215907120.

Park T, Casella G (2008) The Bayesian Lasso. *Journal of the American Statistical Association*, **103**, 681–686.

Pinheiro J, Bates DM (1996) Unconstrained parameterizations for variance-covariance matrices. *Statistics and Computing*, **6**, 289–296.

Porter JH, Nagy E, Kratz TK, Hanson P, Collins SL, Arzberger P (2009) New eyes on the world: advanced sensors for ecology. *BioScience*, **59**, 385–397.

R Core Team (2023) *R: a language and environment for statistical computing.*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rubinacci S, Hofmeister RJ, Sousa da Mota B, Delaneau O (2023) Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. *Nature Genetics*, **55**, 1088–1090.

Silver D, Hubert T, Schrittwieser J, *et al.* (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, **362**, 1140–1144.

Tabak MA, Norouzzadeh MS, Wolfson DW, *et al.* (2019) Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, **10**, 585–590.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–288.

Tredennick AT, Hooker G, Ellner SP, Adler PB (2021) A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, **102**, e03336.

Ustin SL, Middleton EM (2021) Current and near-term advances in earth observation for ecological applications. *Ecological Processes*, **10**, 1.

Wang G, Sarkar A, Carbonetto P, Stephens M (2020a) A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **82**, 1273–1300.

Wang S, McCormick TH, Leek JT (2020b) Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, **117**, 30266–30275.

Weiss KM (2008) Tilting at quixotic trait loci (QTL): an evolutionary perspective on genetic causation. *Genetics*, **179**, 1741–1756.

Wilmers CC, Nickel B, Bryce CM, Smith JA, Wehat RE, Yovovich V (2015) The golden age of bio-logging:how animal-borne sensors are advancing the frontiers of ecology. *Ecology*, **96**, 1741–1753.

Wolpert D, Macready W (1997) No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, **1**, 67–82.

Wolpert DH (1996) The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, **8**, 1341–1390.

Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM (2013) Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, **14**, 507–515.

Xiong W, Droppo J, Huang X, *et al.* (2016) Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*.

Yoo AB, Jette MA, Grondona M (2003) SLURM: Simple Linux Utility for Resource Management. In: *Job Scheduling Strategies for Parallel Processing* (eds. Feitelson D, Rudolph L, Schwiegelshohn U), pp. 44–60, Springer Berlin Heidelberg, Berlin, Heidelberg.

Zhao WX, Zhou K, Li J, *et al.* (2023) A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhou X, Carbonetto P, Stephens M (2013) Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics*, **9**, e1003264.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.

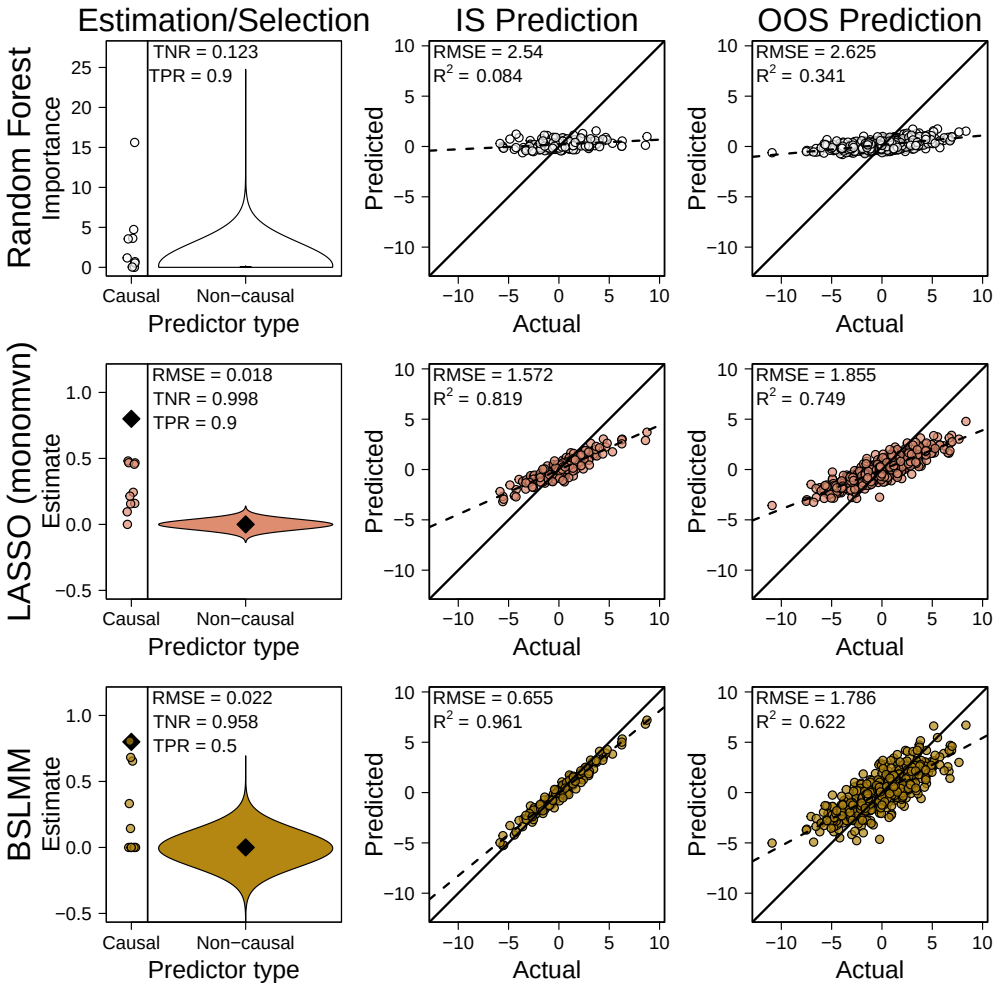

Figure 1: Performance varies greatly across three methods for parameter estimation, variable selection, in-sample (IS) prediction, and out-of-sample (OOS) prediction. Results are shown for the first replicate of scenario 24, which had 10 causal predictors ($\boldsymbol{\beta} = 0.8$) and 9,990 non-causal predictors ($\boldsymbol{\beta} = 0$). The distributions of causal and non-causal importance values are shown for Random Forest, whereas the distributions of causal and non-causal parameter estimates are shown for LASSO `monomvn` and BSLMM (black diamonds signify the true effect sizes). In-sample prediction was based on 150 observations used to train the model, and out-of-sample prediction was based on a separate 500 observations (the maximum reducible error for this scenario was $R^2 = 0.832$). RMSE: root mean square error; TNR: true negative rate (i.e., specificity); TPR: true positive rate (i.e., sensitivity)

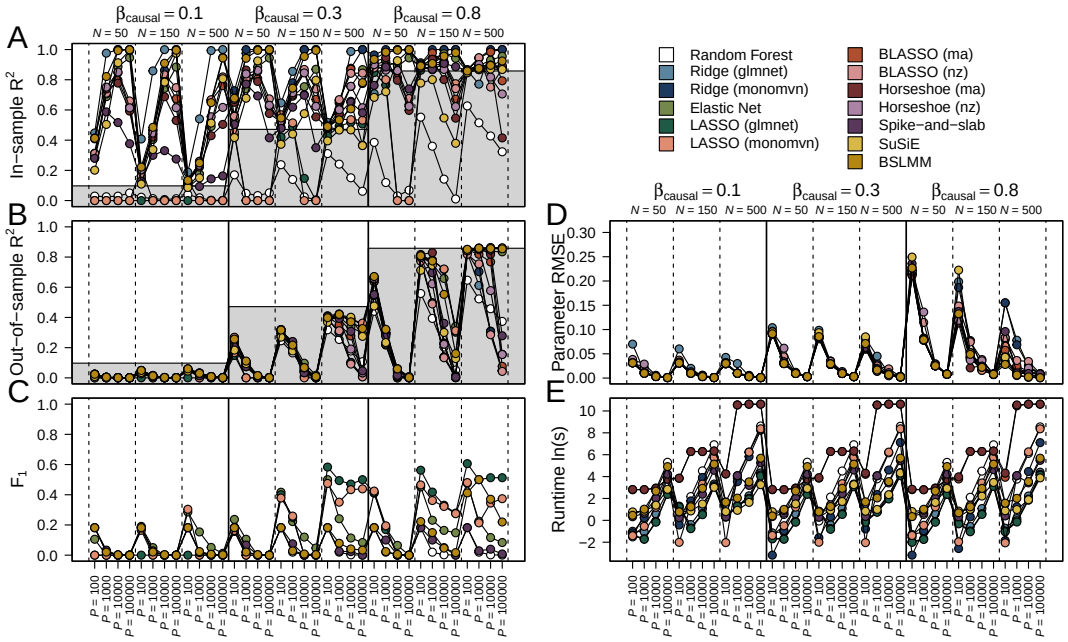

Figure 2: An overview of model performance for the 36 core scenarios. Nine methods were considered, as well as two different implementations of Ridge and LASSO in the `glmnet` and `monomvn` packages in `R`. Metrics for BLASSO and Horseshoe were calculated two ways: model-averaged (ma) estimates are based on all samples from the reversible jump MCMC, whereas non-zero (nz) estimates use only samples in which the predictor and associated coefficient were included in the model. (A) In-sample and (B) out-of-sample prediction were evaluated with R^2 between the actual and predicted values of the response. In these panels, the grey and white regions represent the mean reducible and irreducible error, respectively, across all scenarios within a $\boldsymbol{\beta}_{\text{causal}}$ level. While reducible error represents the expected maximum value for out-of-sample prediction, in-sample prediction can exceed the reducible error when too flexible models are employed (i.e., overfitting). This means that the target for prediction is to recover a model with in-sample and out-of-sample R^2 equal to the maximum reducible error. (C) Variable selection was evaluated using F_1, which is the harmonic mean of precision (i.e., the fraction of selected predictors that are truly causal) and sensitivity (i.e., true positive rate): $\frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity}$. F_1 was only calculated for analyses that can return truly sparse parameter estimates (i.e., $\boldsymbol{\beta} = 0$; BSLMM, Elastic Net, LASSO, Spike-and-slab) or importance values (Random Forest). (D) Parameter estimation was evaluated for all methods except Random Forest using the root mean square error (RMSE) between estimated and actual parameter values. (E) Model speed was evaluated based on the natural log of runtime in seconds. Each circle represents the median value from 100 replicate simulations.

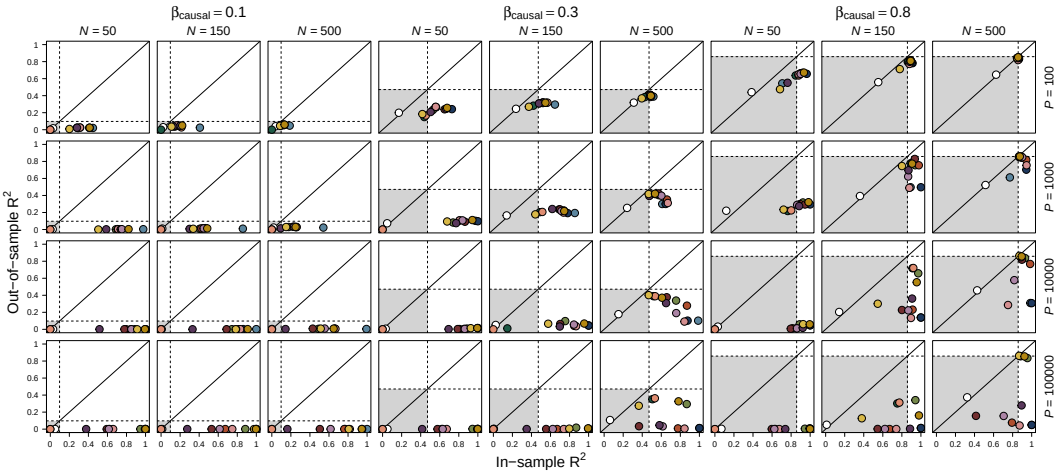

Figure 3: The extent of overfitting and the fraction of reducible error recovered differs dramatically among methods and data attributes. The grey and white regions represent the mean reducible and irreducible error, respectively, across all scenarios within a $\boldsymbol{\beta}_{\text{causal}}$ level. While reducible error represents the expected maximum value for out-of-sample prediction, in-sample prediction R^2 can exceed the reducible error when too sensitive models are employed (i.e., overfitting). This means that the target for prediction is to recover a model with in-sample and out-of-sample R^2 equal to the maximum reducible error, as was the case for many of the methods in the upper right hand panel ($\boldsymbol{\beta}_{\text{causal}} = 0.8$; $N = 500$; $P = 100$). Each circle represents the median value from 100 replicate simulations. See Fig. 2 for color legend.

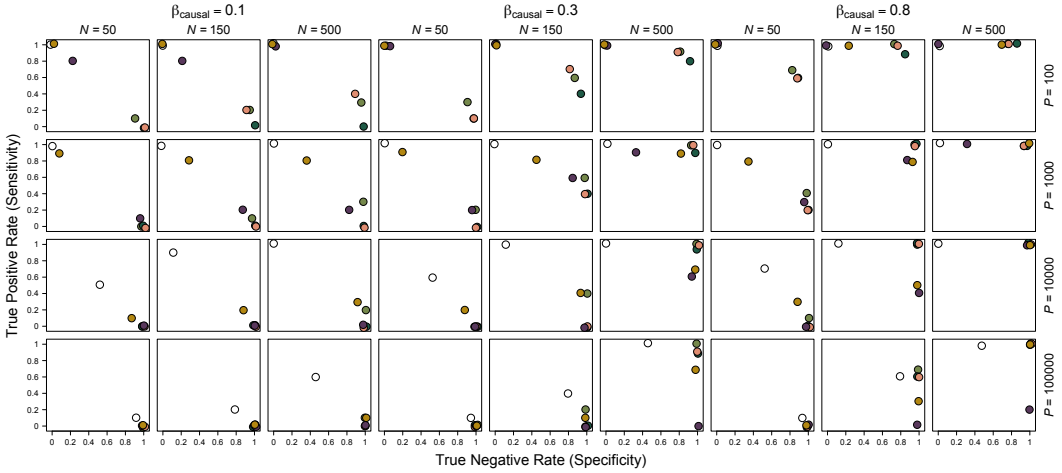

Figure 4: Variable selection performance varied greatly across scenarios, and was only possible for some methods when there were many observations (N), few predictors (P), and effect sizes ($\boldsymbol{\beta}_{\text{causal}}$) were large. A negative correlation between true positive rate and true negative rate emerged for many simulations, especially when $\boldsymbol{\beta}_{\text{causal}}$ was small, indicative of a trade-off between identifying causal predictors (sensitivity) and excluding non-causal predictors (specificity). This trade-off disappears when conditions are more favorable for variable selection: when $\boldsymbol{\beta}_{\text{causal}}$ and N are large and when P is small. Variable selection was only evaluated for analyses that can return truly sparse parameter estimates (i.e., $\boldsymbol{\beta} = 0$; BSLMM, Elastic Net, LASSO, Spike-and-slab) or importance values (Random Forest). Each circle represents the jittered median value from 100 replicate simulations. See Fig. 2 for color legend.

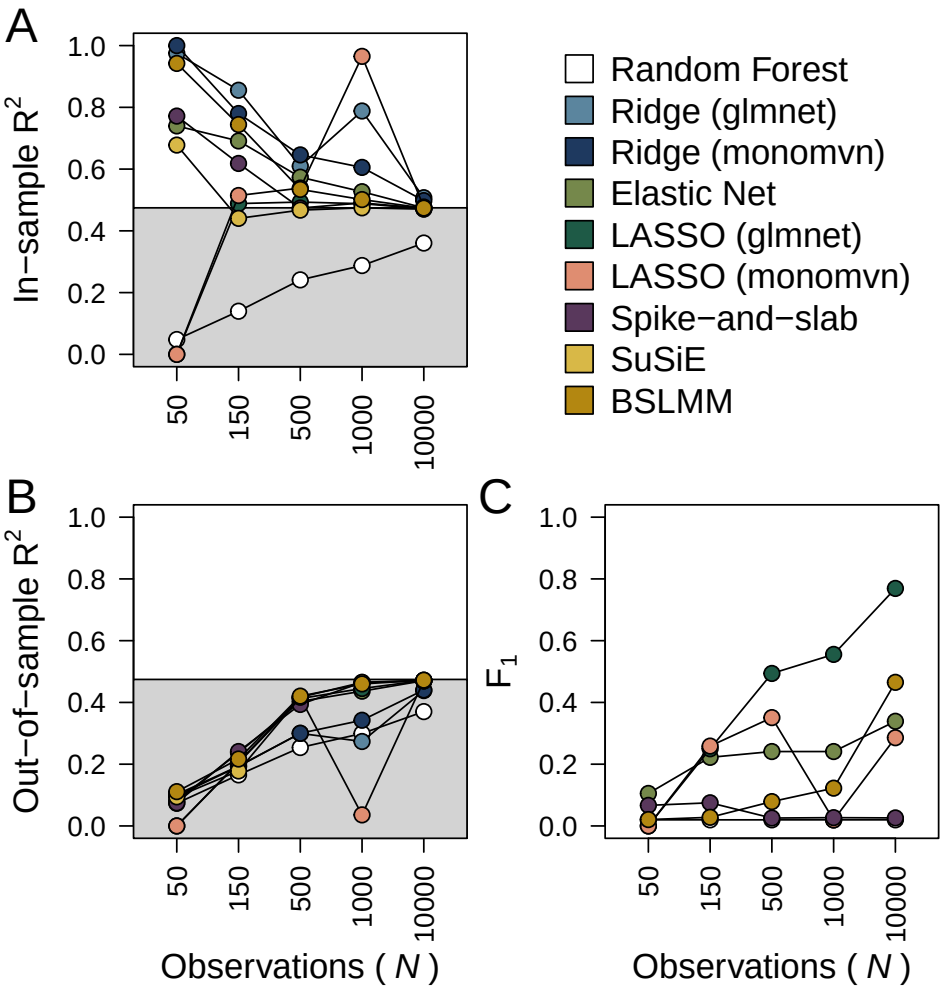

Figure 5: Model performance for (A) in-sample prediction, (B) out-of-sample prediction, and (C) variable selection improves with increasing observations (N) for most models (BLASSO and Horseshoe were not considered because of long runtimes). The five scenarios shown here all had P = 1,000 and $\boldsymbol{\beta} = 0.3$. In-sample and out-of-sample prediction were evaluated with R^2 between the actual and predicted values of the response. In panels A and B, the grey and white regions represent the mean reducible and irreducible error, respectively, across all five scenarios. While reducible error represents the expected maximum value for out-of-sample prediction, in-sample prediction can exceed the reducible error when too flexible models are employed (i.e., overfitting). This means that the target for prediction is to recover a model with in-sample and out-of-sample R^2 equal to the maximum reducible error. Variable selection was evaluated using F_1, which is the harmonic mean of precision (i.e., the fraction of selected predictors that are truly causal) and sensitivity (i.e., true positive rate): $\frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity}$. F_1 was only calculated for analyses that can return truly sparse parameter estimates (i.e., $\boldsymbol{\beta} = 0$; BSLMM, Elastic Net, LASSO, Spike-and-slab) or importance values (Random Forest). Each circle represents the median value from 100 replicate simulations.